WeiboStand: Capturing Chinese Breaking News Using Weibo "Tweets" *

Cheng Fu
University of Maryland
College Park, MD
cfu@umd.edu

Hanan Samet
University of Maryland
College Park, MD
hjs@cs.umd.edu

Jagan Sankaranarayanan
NEC Labs America
Cupertino, CA
jagan@nec-labs.com

ABSTRACT

Weibo is the premier microblog service in China, which is nicknamed as the "Chinese Twitter". Weibo messages consist of text messages, short links, images, audio and video. Its text is restricted to 140 Chinese characters. Since Twitter is blocked in the mainland of China, Weibo is the dominant microblog service in China. The dominance of Weibo in China makes it an obvious choice for capturing late breaking news. This paper describes the implementation of a system for capturing messages corresponding to late breaking news as well as a visualization tool that can display Weibo news messages on a map interface. There are several technical challenges to building this system. First, methods to automatically recognize and disambiguate geographical locations in messages written in Chinese. Second, due to the lack of a free accessible real-time streaming API as that similar to the Twitter Public Streaming API, a new strategy to collect the most recent news-related Weibo messages is devised. The system also uses news from Chinese news RSS feeds as complementary sources.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

General Terms

Algorithms, Design

Keywords

Weibo, News, Geotagging

* This work was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-12-19023, and IIS-13-20791.

ACM SIGSPATIAL LBSN'14, November 4, 2014, Dallas, TX, USA (c) 2014 ACM ISBN 978-1-4503-3140-1...$15.00

1. INTRODUCTION

Weibo is a popular microblog service in China with 500 million registered users and 46 million active users. A recent China-wide investigation shows that 43.6% of the people who have Internet access use a microblog service [3]. Each text message is limited to 140 Chinese characters with additional metadata such as posting and reposting time, count, etc. There has been considerable interest in building news systems that can automatically extract late breaking news events from microblogs and associate them with the location where the news event is happening. One example of such a system is TwitterStand [21], which captures late breaking news using Twitter tweets. Since TwitterStand only works for Tweets written in English, there is a need for a similar system for Chinese users in mainland China where the only predominant microblog service is Weibo and the messages are predominantly in Chinese.

In this paper, we discuss the design of a news gathering system called WeiboStand that automatically extracts toponyms (i.e., references to locations) from plain messages and then displays them on a map query interface by geocoding the toponyms. The system mainly follows the infrastructure design of NewsStand [12, 13,16,19,20,23] and its analogous tweet-based system, TwitterStand [8,21]. It is motivated by our work in browsing spatial data [5,17,18]. WeiboStand accommodates for the difference in natural language processing between English and Chinese microblog messages and adjusting for the differences between Twitter and Weibo.

Building a microblog based newsgathering system for Chinese messages is challenging due to several reasons. First, there are several intrinsic differences between processing English versus Chinese messages. A key difference between a message written in English and Chinese is the expressiveness of the Chinese characters, which can use much shorter length messages than English to convey the same information [14]. A 140-character Weibo message can describe a whole news story with a full paragraph rather than one to two sentences as in the case of English language tweets. For example, the following Weibo message posted by the United Nations has 110 characters, including Chinese characters, numbers and punctuation symbols:

［可持续发展］在太平洋岛国萨摩亚举行的第三届小岛屿发展中国家国际会议 9 月 4 日落下帷幕。各国政府以及非政府组织和私营企业在此次会议上达成了近 300 项合作伙伴关系，承诺为小岛国实现可持续发展提供的资金总额高达 190 多亿美元。

This message translated to English reads as follows: *The 3rd International Conference on Small Island Developing States is finished on September 4th in the Pacific island nation Samoa. Governments, NGOs and private companies achieved nearly 300 cooperation relationships. They promised to offer 19 billion dollars to support the small island states for sustainable development.*

Since users can obtain almost the complete news story from a single Weibo message, Weibo has been become the dominant news source for users in China. Processing Chinese messages is much more complex given the large number of alphabets in Chinese language. Moreover, the Chinese language is constantly evolving in terms of developing new words and new ways of referring to ideas, people, and locations. This is quite different from the rather slow evolving nature of the English language. Next, segmenting a sentence into words or deciding if a phrase refers to a geographical location is more complex to parse given that there is really no demarcating space between words in Chinese. Finally, there are technical issues in using Weibo, which is a much more of a closed service than Twitter. This makes collecting messages a challenge. The lack of good APIs makes building a system on top of Weibo all the more complex, which is the reason we augment messages with RSS news sources to ensure that our system captures sufficient news topics.

The rest of the paper is organized as follows. We first provide an overview of the Weibo service in Section 2 followed by a brief overview of TwitterStand. We discuss issues in segmenting Chinese sentences in Section 4, while issues in toponyms recognition and disambiguation are reviewed in Section 5. We present the system architecture of WeiboStand in Section 6 and draw concluding remarks in Section 7.

2. WEIBO SERVICE

Weibo is popularly known as "Chinese Twitter" since it virtually provides a nearly identical service to Twitter except that the messages are in Chinese. Registered users can post a short message limited to 140 Chinese characters. Each message contains a main body text, a time stamp, a unique serial number, repost information, and URLs for images and videos, etc. Users can follow other users to get their most recent messages. However, not all messages are sent to all followers. In this respect, Weibo is similar to Google Circle in the sense that users can specify which subset of their followers should receive a message.

Weibo offers clients on webpages, mobile phones and tablets. Weibo also provides APIs for developers to build apps. However, Weibo has more restricted control on data and privacy. Developers must submit an application to Weibo with a full description of the purpose of their apps. For unauthorized apps, there are even more restrictive limitations [25]. Unlike Twitter, Weibo does not provide a sampled stream as an API to developers similar to those provided by Twitter (e.g., the GardenHose which is a 10% sampling of all tweets in Twitter). Furthermore, a recent policy change in September 2013 has effectively banned third-party apps from collecting and storing users' messages through the API without authorization. Weibo makes its messages unsearchable by search engines so no search engine service currently indexes Weibo messages. Therefore, it is important to design a strategy that can fetch the most recent news Weibo messages without relying on Weibo API.

Weibo also provides an identification service so that people can get official verification of their accounts. To be verified, organizations, individuals and companies submit an application to Weibo that includes copies of an official identification document. Once verified, Weibo will put an endorsement badge on these accounts. Many international organizations, news organizations, government officials and journalists have verified accounts. Some of the verified accounts include the United Nations (http://weibo.com/un), the Phoenix News (http://weibo.com/phoenixnewsmedia), the Embassy of the United States (http://weibo.com/usembassy), etc. This verification mechanism is important since we can use it to collect reliable breaking news reports.

Another feature of Weibo is a messages board where messages are organized by topic. News is contributed by users are collected and ranked by Weibo. This feature is similar to Google News in the sense that it provides a wide spectrum of news drawn from different users. One missing feature in Weibo is a service that identifies *where* events are happening. Although most news has location information to identify where the event happens, it requires readers to have a rich knowledge of geography to be able to associate news topics with its location of interest. To answer the question of where a news event is happening, we need a system with a map query interface where the news messages are positioned at their location of focus.

Each Weibo user has a unique serial number and their own profile webpage with a unique URL to show their messages ordered from the most recent to the oldest. Without logging in, users can still access other user's personal profile webpages via a URL and read their most recent 10 public messages. Since Weibo has no mechanism such the Twitter Public Streaming API allowing developers to access a sample of messages in real-time, we will rely on this feature to gather messages from users. In other words, we will periodically crawl the recent messages from select users and collect their last 10 messages

3. TwitterStand Overview

TwitterStand [21] is a news processing system that collects tweets, selects news-related tweets, aggregate similar news tweets, geotags news tweets and displaces news tweets on map. TwitterStand utilizes Twitter's GardenHose and BirdDog as its main tweet sources. GardenHose offers 10% sampling of tweets while BirdDog provides tweets from up to 200,000 users. It also uses a set of "seeders" that correspond to users or organizations

that predominantly tweet about news. Furthermore, TwitterStand also uses the Twitter search API to continuously gather tweets by keywords or hashtags.

After collecting new tweets, TwitterStand separates out the tweets that may correspond to news. It then extracts location information from the text of the tweets, their geo-tags or the location of the accounts. The system groups tweets with similar information. Next, it posts the tweets on a world map to show where the events are happening.

One major limitation of the TwitterStand system is that it is limited to tweets in English. Therefore, it can catch breaking news if enough number of people express the news in English. This is not true for local news events in China.

4. SEGMENTATION OF SENTENCES

One significant difference between Chinese and English is that Chinese sentences do not use spaces to demarcate between words. For instance, 我在马里兰。 (I'm in Maryland) has no space symbols to segment the sentence into words. Therefore, the first step of natural language processing for Chinese sentences is to breakup a sentence into words correctly using a segmenter.

Designing a segmenter for Chinese is not an easy task. Unlike English letters, each Chinese character has its own meaning. Most of them can be used independently. For instance, in the case of 我在马里兰。, the five characters mean *I*, *In*, *Horse*, *Inside* and *Orchid*, respectively, while the sentence in its entirety should be read as "I am in Maryland." Characters can also be combined into words or phrases. Furthermore, single characters, words, and phrases are combined into a sentence. Since Chinese sentences do not use space to separate characters, the phrase 结婚的和尚未结婚的 (married and unmarried people) can be separated into its correct components (i.e., 结 婚 的 / 和 / 尚 未 结 婚 的 married/and/unmarried. An incorrect segmentation of the sentence is 结婚的/和尚/未结婚的 (married/monk/unmarried). Even though the latter solution splits the phrase incorrectly, the three constituent words are still meaningful independently.

It is easy to create new words in Chinese as any combination of Chinese characters can be used as a new word. The only caveat is that the use of word should be popular. Interestingly, Weibo due to its popularity is actually creating many new Chinese words. Weibo (微博) itself is a new word. Other examples include 土豪 (rich but not-well-educated people), 团购 (groupon), 雷人 (astonishing), etc. are examples of new words created by users of Weibo. These words are not only used by social media users but also are finding use in conventional media. However, this also poses a challenge for Chinese NLP packages since they need to constantly keep up with the new words that are constantly created and used in Weibo messages.

5. TOPONYMS IN CHINESE

Once a sentence has been segmented, we need to identify those words in the sentence that are references to geographical locations (i.e., toponyms). In this section, we discuss the issues in performing toponyms recognition and resolution in Chinese language.

5.1 Toponym Forms

Toponyms in Chinese have three forms, which we call full name, abbreviation and single character.

Full name toponyms are usually the official and formal name. For instance, 美利坚合众国 is the official name of the United States of America in Chinese; 北京市 corresponds to the city of Beijing. However, as the full name is usually long, they usually appear only in official statements or at the very beginning of the news article or tweet.

Abbreviation toponyms are the short versions of a full name toponym. For instance, 美国 (America) is short for 美利坚合众国 (the United States); 中国 (China) is short for 中华人民共和国 (the People's Republic of China); 俄罗斯 (Russia) is short for 俄罗斯联邦 (the Russian Federation). Abbreviations are common in news and other informal documents.

As abbreviations are still long, single-character toponyms are also common in news. Single characters toponyms use only one Chinese character to refer to a place. Usually, this single character is from part of the full name. For instance, 美 stands for 美国 (U.S.); 俄 stands for 俄罗斯 (Russia); 中 or 华 stands for 中华 (alias of 中国 China); 京 stands for 北京 (Beijing). However, for historical reasons, many Chinese cities have their own single-character aliases, which are not part of their full name. For instance, 沪 is the single-character name of 上海 (Shanghai). It is the only single-character name of Shanghai. Neither 上 nor 海, which is part of the full name of Shanghai, stands for Shanghai.

The problem of single-character toponyms is that the same character referring to a place can also refer to very different meanings as one Chinese character usually has several meanings depending on the context. For instance, 美 stands for the United States also has means *beautiful* and *beauty*. 日 stands for *Japan* as well as means *sun, date, day time* or *schedule.* 法 stands for France, but also means *law, principle,* or *mimic*.

For this reason, although single-character toponyms make up roughly 34% of overall abbreviations in a newspaper, segmenter algorithms have low precision in the task of extracting single-character toponyms from sentences [9].

5.2 Disambiguating Toponyms and Person Names

One peculiarity of Chinese language processing, which it shares with the English language is that the same characters can identify both a person and a toponym. . The number of different last and first names in English is relatively small and they are capitalized in sentences. Han Chinese names usually consist of a last name consisting of one character followed by a first name consisting of no more than two characters. There is no limitation on the characters that make up a given name. A source of ambiguity is

that most Chinese cities also use two characters, which means that it easy to mistake a Han Chinese name for a toponym.

To avoid misrecognizing a name as a toponym, a segmenter usually maintains a list of famous people who often appear in news. However, this approach is hardly robust and fails for even simple cases. For instance, the correct segmentation of 洪贝宁在贝宁约见撒贝宁。 (Hong Beining meets Sa Beining in Benin) is 洪/贝宁/在/贝宁/约见/撒贝宁/。. Here, 洪贝宁 and 撒贝宁 are persons while 贝宁 is the abbreviation for 贝宁共和国 the Republic of Benin. A segmenter correctly segments 撒贝宁 since he is a famous TV anchor in China and is present in the segmenter's dictionary, while Hong Beining 洪贝宁 who has the same given name as Sa Beining but a different last name is mistakenly segmented into 洪, which has no meaning in this context, and 贝宁, which is Hong's first name but also the same as the Chinese abbreviation name of the Republic of Beinin.

Another common source of error is when a place is named after some famous person's given name. As most traditional Chinese city names use two characters, given names are used to name a city unless the person's full name has only two characters. For example, 中山 is a city in Guangdong Province named after 孙中山 (Sun Yat-sen). 靖宇 is a county in Jilin Province named after 杨靖宇 (Yang Jingyu). Of course, this is also a common problem in English where Washington could be a name of a person or could be a toponym depending on how it is used in a sentence (e.g., [10]).

Non-Chinese names can be written in multiple ways in Chinese. For instance, names can be written as spelled in their original language (e.g., English) and can be transliterated into Chinese characters. In news, transliterated Chinese written names are commonly used. These names are usually long and could possibly be missing in the segmenter's dictionary since there is no real standardization. For instance, 麦迪娜·买买提 (Madina Memet, a Uyghur actress), 吉克隽逸 (Jike Junyi, a Yi pop singer), etc. require several characters to represent in Chinese, similarly for other foreign names that are also transliterated into Chinese characters, e.g. 奥巴马 (Obama) and 内塔尼亚胡 (Netanyahu). For Japanese and Korean names, they are used as what they are in form of Japanese Chinese characters and Korean Chinese characters, e.g. 潘基文 (Ban Ki-moon) and 安倍晋三 (Abe Shinzo). Since these non-Chinese names are likely not in the segmenter's dictionary unless they are famous, they could be incorrectly split and can incorrectly match with toponyms.

5.3 Toponym Ambiguity

The administrative division naming system in China follows the following hierarchy: *province, prefecture, county/district, town* and *village*. Duplicate toponyms in China do not arise as frequently as in the United States and other countries. According to the 1986 Toponym Management Regulations [22] of China, no foreign toponyms or names should be used as toponyms; no administrative sub-divisions within the same administrative division should share the same name; no *important* toponyms within a province should share the same name.

Ambiguity in toponyms mainly arises in three ways. First, a prefecture shares the same name with one of its counties only if the prefectural government is located in that county, similar to the case of New York City and New York State. Second, a common source of ambiguity happens when places are referred to by their relative position to major cities. For instance, all major cities in China have districts that are commonly referred to by their relative position with respect to a major city, e.g. 东区 (East District) and 城北区 (North District of the City). Third, a district can have the same name as another district or town or village outside of its own prefecture [7], which is also a common source of ambiguity in other countries such as the United States where there are multiple counties with the same name in different US states.

In English, containment relationship of places appears as comma separated pairs in text. For example, the reference 'College Park, Maryland' leads us to infer that the place *College Park* is within the place *Maryland.* This is an easy pattern to identify. However, in Chinese, place names are written in an order so that the larger geographical unit appears first with smaller geographic units following with no comma separators. Therefore, in Chinese, such containment relationships are difficult to identify if the toponyms are embedded in sentences.

Although the task of disambiguating toponyms in Chinese is not as hard as that in English in terms of the number of different occurrences of duplicate toponyms (e.g., [11,13,15]), the challenge are different as we have outlined.

6. ARCHITECTURE

WeiboStand's architecture is a simplified version of TwitterStand. Since there are no equivalent API services as GardenHose, or BirdDog, the seeder mechanism of TwitterStand is retained in WeiboStand. To create the seeders feed, we manually identified a few hundred accounts of users in Weibo who mainly send messages about breaking news. In particular, these correspond to journalists, bloggers, news agencies, government officials etc. WeiboStand periodically crawls their profile age and extracts the recent messages. These messages are processed following the same mechanism as in TwitterStand, although one major difference between Chinese and English message processing is that segmentation of sentence into words is an additional yet critical module in WeiboStand.

Figure 1 provides the architecture of WeiboStand. The system consists of data collection, processing and displaying modules. The data collection module is composed of an RSS news crawler and a Weibo crawler. The RSS crawler regularly downloads 200 online news sources and updates a database of news articles. The Weibo crawler monitors the latest messages of 300 seeders and stores the new messages into the database.

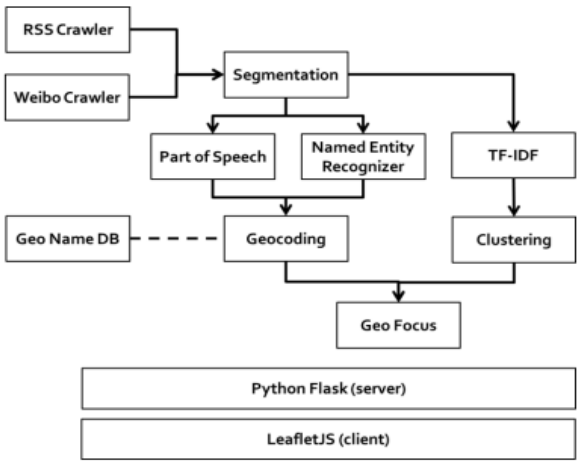

Figure 1 Architecture of WeiboStand

Next, each news message is passed to the named entity recognition (NER) or part of speech labeling (POS). As Chinese character sequences are segmented in different ways, POS and NER implicitly perform the segmentation step before POS labeling of a sentence is performed. WeiboStand uses the Stanford Chinese NLP package. The Stanford Chinese NLP package has an F-score of 0.95 on the Peking University corpora [24], and also has a high score in recognizing "out-of-vocabulary" words [2]. To our best knowledge, there is very little systematic comparison of the performance of segmenters. One possible reason is that the result may vary based on corpora and standards. The Stanford segmenter is based on conditional random fields (CRF) [24]. It uses several updated Chinese dictionaries, which is important for Weibo messages where new words can be created often. It is also constantly updated as new words are created. For instance, the last revision of the Stanford NLP package was on Aug. 2014. There are other segmenters available for Chinese language but they are not updated as often, which makes them not suitable for segmenting Weibo messages. For instance, the last update of ICTCLAS by the Chinese Academy of Science was released in 2012. Another benefit of the Stanford NLP package is that it also comes with an NER and a POS tagger.

To extract toponyms, NER and POS are used. NER can identify words by entity classes, such as PERSON, ORGANIZATION and LOCATION. The words labeled as LOCATION are the most likely toponyms. However, we often found that toponyms are often misclassified as ORGANIZATION so we also treat these words as toponym candidates. Furthermore, we also consider words labeled as proper nouns by POS as toponym candidates. Since the gazetteer extracted from GeoNames database has no single-character toponyms, the candidates with only one Chinese character are removed from the toponym candidate set.

Next, the toponym candidates are processed by the geocoding module, which finds the corresponding coordinates of the candidates in a database that stores name-coordinate records.

The messages are then clustered into a news articles using a document-clustering algorithm. This is done by extracting word features by Term Frequency-Invers Document Frequency (TF-IDF) and clustering word features according to their similarity in the clustering module.

Once clustered, WeiboStand finds the geographical focus of each news cluster. Each toponym is assigned a geographic coordinate value corresponding to its most common interpretation in the messages belonging to the cluster. The geographic focus is determined by performing a clustering algorithm that determines a few geographical areas that are mentioned in a majority of the messages within the cluster.

A web client is implemented for visualizing the geocoded clusters as shown in Figure 2. In the figure, we show several news stories on a map query interface. They correspond to news obtained from Weibo or from RSS feeds. We mark the messages obtained from Weibo using an "eye" marker (Sina Weibo's logo) and those obtained from RSS feeds using a blue marker. By clicking the marker, the news story pops up with the toponym being highlighted. Our online demo is available from http://weibostand.umiacs.umd.edu/.

Figure 2 Web interface of WeiboStand. Weibo logos (eye shaped marker) show news clusters derived from Weibo. Blue markers show news clusters derived from RSS news sources.

6.1 Chinese Gazetteer Preparation

A gazetteer in Chinese is critical to figure out the coordinate values corresponding to a toponym. The Chinese gazetteer in this system is extracted from the GeoNames database (http://www.geonames.org/). GeoNames is a free geographical database providing a worldwide name-coordinate lookup service. However, the main entry of the database is English-oriented. Only parts of its records have Chinese character aliases in UTF-8. A preprocessing module was implemented to extract the UTF-8 Chinese alias-coordinates pairs from the original GeoNames database and store them into our own local gazetteer database. A highlighted issue is that the UTF-8 Chinese character set is larger than the Chinese characters used by Chinese. Chinese characters are used in China, Japan (汉字, *Kanji*) and Korea (한자, *Hanja*). Japanese and Korean use both original Chinese characters and the Chinese characters created by Japanese and Korean. For historical reasons, there are also simplified Chinese characters (used in

mainland China and Singapore) and traditional Chinese characters (used in Hongkong, Macau and Taiwan ROC). These four forms of Chinese characters share some characters. In UTF-8, if a character has the same shape in these four Chinese character sets, then it has only one code, e.g. 日 *sun* has one code (U+65e5), as it has the same shape in four forms. 對 *right* (U+5c0d) and 对 *right* (U+5bf9) have different codes in UTF-8, but they are actually the same character in traditional form and simplified form. Because of this issue, the Chinese gazetteer we derived contains not only Chinese toponyms but also Japanese toponyms in the form of Kanji.

Our Chinese gazetteer that is derived from GeoNames contains 337,297 records. As Japanese cities are officially written in Japanese Chinese characters (Kanji) rather than Japanese characters (假名, *Kana*), Japanese toponyms take a large proportion of our gazetteer. Table 1 shows the distribution of these toponyms and we see that most of the places are within China (including mainland China, Taiwan ROC, Hong Kong and Macau).

Table Top 10 Countries with places in Chinese in gazetteer

Country	Num.	Country	Num.
China	291553	Poland	475
Japan	24693	UK	351
US	2622	Spain	344
Germany	1408	Russia	300
France	1335	India	269

Although GeoNames has a professional team maintaining the original gazetteer, we observed several issues with it. First, a toponym in GeoNames can have multiple coordinate values. For instance, there are 8 records for 北京 *Beijing*, which include a reference to the centroid of city of Beijing, Beijing's administrative region, and urban area of the city of Beijing, etc. The second issue is that the toponyms in Chinese within this gazetteer are neither evenly distributed in space or in a spatial hierarchy. Many toponyms for places outside of China are missing, e.g. there are only 2611 toponyms for the US. However, there are also small places are in the gazetteer. For instance, the Burj Khalifa Tower of Dubai is in this gazetteer. The third issue is that places are presented as points in this database even though they may have extents. This makes it difficult to organize these toponyms into a spatial hierarchy, which can improve *toponym resolution*. One heuristic resolution proposed in [23] is to use the spatial hierarchy. For instance, in the phrase 'London, Canada', there is a 'London' and 'Canada'. This phrase most likely refers to a 'London' *in* a 'Canada'. If such a relation can be found in the database, then the system should return a single record rather than several records referring to several places named London. This requires prior knowledge about the spatial relationships among the records in the gazetteer. We perform a preprocessing step in order to fix these three issues before we use it in our system.

6.2 Weibo Message Crawling

For Weibo messages, we periodically crawl the webpages of our seeders. Our seeders are those users whom we have pre-selected as reliable providers of current news (e.g., [6]). They include domestic newspapers, international newspapers, international organizations, foreign embassies, journalists, government officials and the embassy accounts offering local information. Since Weibo has been prevalent as an online news source for people, most traditional news sources, including newspaper presses, national and local TV stations, news websites, local channels, journalists and government agencies of foreign countries, have Weibo accounts. Even though the news messages from these accounts are duplicates of the news in newspapers and TV stations, the messages can spread much faster than conventional news media. Weibo and other websites maintain their own local channel accounts that post localized daily information and news. These accounts are good sources for the news that are highly localized and beyond the radar of conventional media. Accounts of international organizations and foreign government agencies are good sources of international news events as Chinese domestic news media can usually only cover big international news at the national level rather than local events in foreign countries. The number of such seeders is limited, which then limits the number of Weibo messages that can be collected, but messages from these accounts can all be treated as news messages as they are reliable news sources.

For third-party sources, the most recent 10 messages can be found in the profile page of a user without having to log on. The crawler script rolls over the seeder list, refreshes and parses their profile pages to extract these public free accessible messages. If a message ID is not in the database, then the script will identify it as new and enter it into the database.

Weibo has different web page structures for different account groups (with or without identification; commercial partners or regular users). Three different web page templates are identified after observation. Each template uses a different HTML element structure to wrap messages. There is no official document to explain the exact structure. Fortunately, each template has a flag in its HTML file. Thus, the message content and critical metadata, including post time and the unique ID, can be extracted via DOM querying packages.

6.3 RSS News Crawling

Even though news from social media like Weibo has been a major news source for Chinese people, Really Simple Syndication (RSS) is also an important complementary news source. RSS news sources are more reliable since conventional media publishes them. RSS is a widely used XML protocol. The recent RSS 2.0 has rich metadata including title, description, link to original news webpage, and a publication date. Once a new RSS feed comes, a crawler script can download the entire article by following the link and storing the article into the database. The advantage of RSS feeds over Weibo messages is that the articles can cover a topic with more details.

We maintain a list of RSS news sources, including national and local news presses, local broadcast channels in new websites, and the Chinese broadcast channels of foreign news media.

6.4 Geocoding

As mentioned before, the geocoding module receives a list of candidate words that might be a toponym. Each candidate then issues a query to the gazetteer to check if the word exists in the gazetteer. The gazetteer returns a list of coordinate pairs of latitude and longitude that match the word in the gazetteer. We calculate the centroid of the list of coordinate values and choose the nearest coordinate value to the centroid in the list as the resolution of the toponym. The rationale for this heuristic approach is that a place with several records spatially clustered in the gazetteer may be a place with a higher administrative hierarchy and more famous. This makes it be more possible as the correct target place instead of other records that share the same name but are located farther away. In the case of Beijing, the 8 records of Beijing are clustered around the *real* Beijing. Their centroid should also be within the realm of Beijing and thus, the nearest record to this centroid should be in Beijing. Even if there was another small Beijing located in another province, the median centroid can still ensure that the one of the cluster members is returned.

6.5 Geographical Focus

Each message may contain several toponyms that all have corresponding coordinates in the gazetteer. This is referred to as the geographical focus problem [23]. To determine which location is the focus of the news message, we adopt the strategy used in NewsStand. We assume that the closer the relative position of a toponym is to the beginning of a news message, the more likely that it is the geographical focus of this message. In the meantime, if a toponym appears several times in one message, then this toponym could be the geographical focus. To balance these two situations, each toponym is assigned an index to identify its position in the message. For RSS messages, title is placed before the body, because the toponym in the title may be the most likely location referred to by the message. Then for each toponym, the sum of the inverses of the position index values is used as the score. The toponym with the highest score is assigned as the geographical focus of this message.

After the messages that cover the same event have been clustered, it is also necessary to determine the geographical focus of the cluster. Majority vote is used in our system because we assume that the messages describing the same event should have the same geographical focus.

6.6 Clustering

Since the messages and articles may cover the same story, it is necessary to group relative messages and articles together. These groups are called *clusters*. Since messages from Weibo are no longer than 140 Chinese characters and articles from RSS sources usually contain as many as several thousand characters, we cluster messages from Weibo and articles from RSS separately. .

6.6.1 Preprocessing

Once a news item from a message or an article is clustered, we first segment it into a sequence of words. We then extract the news item's term feature vector by computing its TF-IDF score.

The TF-IDF score emphasizes the words, which are frequent in particular documents and infrequent in the overall corpus. The TF-IDF score of a word *i* in an article *j* is:

$$TF - IDF_{i,j} = \frac{n_{i,j}}{n_j} log^{\frac{|D|}{o_j}}$$

where $n_{i,j}$ is the frequency of word *i* in document j, n_j is the number of words in j, $|D|$ is the number of articles, o_i is the number of articles contain word *i*. In our system, the new incoming articles are cached until the number of articles exceeds a threshold or a time threshold is reached. In this way, the TF-IDF score calculation task does not need to process all the collected articles. Several experiments show that it does not significantly affect the clustering result

6.6.2 Clustering

The clustering approach is based on the similarity of words in two articles and the time difference. The similarity of words is measured by *cosine similarity*. The cosine similarity of an article and an article cluster is defined as:

$$\delta(a, c) = \frac{\overrightarrow{TFV_a}}{||\overrightarrow{TFV_a}||} \frac{\overrightarrow{TFV_c}}{||\overrightarrow{TFV_c}||}$$

where $\overrightarrow{TFV_a}$ is the word vector of article a and $\overrightarrow{TFV_c}$ is the word vector of article cluster *c*.

Concerning the temporal effect on news, time is introduced as a weight of the similarity [20]. The modified cosine similarity is formulated as:

$$\delta(a, c) = \frac{\overrightarrow{TFV_a}}{||\overrightarrow{TFV_a}||} \frac{\overrightarrow{TFV_c}}{||\overrightarrow{TFV_c}||} e^{\frac{-(T_a - T_c)^2}{2\sigma^2}}$$

where T_a is the article's published time, T_c is the time centroid of the cluster *c* and σ is the standard deviation of time in cluster *c*.

In practice, the words ranking in the top 20 TF-IDF scores are picked as the feature vector for an article and a cluster. The comparison between an article and a cluster whose time difference exceeds 72 hours are pruned to improve performance.

6.7 Full-text Search in Database

Once articles and messages are stored, clustered and indexed in a database, full-text search should be employed for querying articles and messages by key words. The full-text search of Chinese articles also involves the same segmentation issue that was discussed in conjunction with the geocoding process. The CRF++ extension for PostgreSQL is used to achieve Chinese full-text search in the database. CRF++ is an open source implementation for word segmentation and provides extensions for databases [4].

7. CONCLUDING REMARKS

We described the implementation of an information visualization system that can present Chinese news message that are crawled

from Weibo, the prevalent Chinese microblog service, in geographical space. The system overcomes the lack of a public API to massively fetch the Weibo service by utilizing a limited number of freely accessible Weibo messages.

The system can help to analyze the type of content in which the Chinese media and people are interested. Comparing the locations of Weibo messages and RSS articles may help to understand the divergent preferences of conventional Chinese media and social media.

The system needs improvement especially on removing ambiguity of locations that have the same name. It should be an important step to improve the accuracy of geotagging messages. Previous work (e.g., [1,23]) has proposed several effective ways to reduce this ambiguity, the majorly based on a containment relationship between different toponyms. However, since GeoNames database does not have topological information of its records and building such Chinese toponyms is time-consuming, we do not implement them in the current system.

Due to the limitation of a restricted access to real-time Weibo messages, a direction for future work could be the implementation of a mobile phone app with the same functionality. Since a mobile phone app with Weibo can get authorization from Weibo users on accessing their historical messages as well as real-time messages, mobile phone clients can help the server to retrieve more Weibo messages to improve the system's performance in terms of the data sources.

8. REFERENCES

[1] I. Bensalem and M. K, Kholladi, Toponym disambiguation by arborescent relationships. *Journal of Computer Science,* 6(6), 653-659. 2010.

[2] P.C. Chang, M. Galley and C.D. Manning. Optimizing Chinese Word Segmentation for Machine Translation Performance. *In StatMT.* pp. 224-232, June 2008.

[3] China Internet Network Information Center. *34th Statistical Report on the Internet Development in China*. Retrieved Sept. 2014 from http://www1.cnnic.cn/IDR/.

[4] CRFF++ http://sourceforge.jp/projects/sfnet_crfpp/ . Retrieved Oct 10th, 2014

[5] C. Esperanca and H. Samet. Experience with SAND/Tcl: a scripting tool for spatial databases. *Journal of Visual Languages and Computing*, 13(2):229-255, Apr. 2002.

[6] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. In LBSN, pp. 44-53, Nov. 2013.

[7] S. T Huang,. 中国县级以上行政区划专名重名一览. *中国方域- 行政区划与地名*. (Translation: Chinese special administrative divisions above the county level were duplicate names list) 1997, volume 1, pp. 8-9.

[8] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *LBSN*, pp. 25-32, Nov. 2011.

[9] B. Li and F. Fang, Single Chinese Character Country Name Recognition. *Computer Engineering and Applications.* 28 167-169, Oct. 2006.

[10] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR'11*, pp. 843-852, July 2011.

[11] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *SIGIR'12*, pp. 731-740, Aug. 2012.

[12] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. *GIS*, pp. 179-188, Nov. 2012.

[13] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pp. 201-212, Mar. 2010.

[14] Z. D. Qu, 从汉英篇幅差异比较汉英语的信息密度 (translation: Compare Information Density of Chinese and English by Textural Length). *Journal of Foreign Languages.* 3, 23-26, May 1998.

[15] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *GIS*, pp. 43-52, Nov. 2010.

[16] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *GIS* pp. 525-528, Nov. 2011.

[17] H. Samet, H. Alborzi, F. Brabec, C. Esperanca, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *CACM*, 46(1):63-66, Jan. 2003.

[18] H. Samet, A. Rosenfeld, C. A. Shaffer, and R. E. Webber. A geographic information system using quadtrees. *Pattern Recognition,* 17(6):647-656, Nov/Dec 1984.

[19] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *CACM*, 57(10):64-77, Oct. 2014.

[20] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *WWW*, pp. 257-260, Mar.-Apr. 2011.

[21] J. Sankaranarayanan, H. Samet, B. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In *GIS*, pp. 42-51, Nov. 2009.

[22] State Council of China. 地名管理条例 (Translated: *Toponym Management Regulations of China)*, 2009.

[23] B. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS*, pp. 144-153, Nov. 2008.

[24] H. Tseng, P. C. Chang, G. Andrew, D. Jurafsky, and C. Manning. A Conditional Random Field Word Segmenter for Sighan Bakeoff *In SIGHAN Workshop on Chinese Language,* pp. 68-171, Oct. 2005

[25] Weibo API. http://open.weibo.com/wiki/API. Retrieved Dec. 2013.